The pitfalls in fitting GARCH processes

Gilles Zumbach [1]

October 16, 2000

Abstract

Optimization of GARCH(1,1) processes by maximum likelihood has been documented to be numerically difficult and prone to error. In this paper, it is shown that this difficulty is related to a one dimensional manifold in the parameter space where the likelihood function is large and almost constant. This manifold is revealed by introducing other natural parameterizations for the process, namely other coordinates in the parameter space. In order to avoid spurious fits, we suggest fixing the average volatility of the process by a moment estimate and estimating the remaining parameters with a maximum likelihood procedure. Moreover, the convergence properties of the maximization algorithm are vastly improved by working with one of the new coordinates system. We also investigate the finite sample distribution of the estimated parameters computed with a maximum likelihood. The results indicate that the intercept parameter α_0 is strongly biased, whereas the volatility and several of the new coordinates have smaller bias. The aggregation properties of GARCH(1,1), from 14 minutes to 8 days are investigated. The results indicate that the theoretical aggregation relations for this process are not consistent with the foreign exchange data. Finally, the change of coordinates is generalized to GARCH(p,q), and the empirical property of the log-likelihood is investigated.

JEL: C22

Keywords: GARCH, log-likelihood, time aggregation, parameter bias

[1] Olsen & Associates, Research Institute for Applied Economics
Seefeldstrasse 233, 8008 Zürich, Switzerland.
E-mail: gilles@olsen.ch, Phone: (+41 1) 386 48 24, Fax: (+41 1) 422 22 82.

1 Introduction

Since their inception, GARCH processes have gained a very fast acceptance in the finance literature. The seminal papers are (Engle, 1982; Engle and Bollerslev, 1986; Bollerslev, 1986), and two comprehensive reviews are (Bollerslev et al., 1992) and (Bera and Higgins, 1993). Garch type processes model the volatility and in particular incorporate the long memory, or clustering, observed in financial data, namely the conditional heteroskedasticity. The GARCH family contains a number of parameters which must be estimated on actual data for empirical applications. The estimation is carried out with a standard log-likelihood maximization, using a conjugate gradient-like algorithm to solve the maximization problem. Possibly, more sophisticated algorithms can be used, combining the efficiency of a local search with the safety of a global search. Yet, it is "common knowledge" among practitioners that the GARCH parameters are numerically difficult to estimate in empirical applications. The numerical algorithm can easily fail, or converge to erratic solutions. Therefore, the resulting fitted parameters must be examined with a healthy dose of scepticism.

In this paper, we investigate why the parameter estimation of the GARCH processes are so difficult numerically. Using a change of coordinates in the parameter space, we show that a one-dimensional manifold of nearly degenerate solutions exists. This family of solutions explains the weakness of the fit with respect to the data quality and to the searching algorithm. In particular, with insufficient data, or with corrupted data, the global solution may well not exist or be meaningless. After exploring this manifold of solutions and its origin, we propose to modify the estimation procedure of the parameters. A first parameter, fixing the volatility of the process, is computed by a moment estimate. In a second step, the remaining parameters are fitted using a log likelihood method, which allows for a robust estimation framework.

This paper is organized as follows: in section 2, the GARCH(1,1) process equations are rewritten in different forms in order to make the properties of the process explicit. The log-likelihood function is analyzed in various parameterizations in section 3. In section 4, the cause for the manifold of nearly degenerate solutions is explained. In section 5, the estimation on finite samples, and in particular the size of the bias, are investigated. Section 6 is concerned with estimates of GARCH(1,1) process at higher frequency, and the comparison with the theoretical aggregation relations. The generalization for the GARCH(p,q) processes is carried out in section 7. Finally, the recommendations for robust estimates are presented in the conclusion.

2 Rewriting the GARCH(1,1) process

The GARCH(1,1) process is usually written as

$$r_i = \sigma_i \varepsilon_i \tag{1}$$

$$\sigma_i^2 = \alpha_0 + \alpha_1 r_{i-1}^2 + \beta_1 \sigma_{i-1}^2 \tag{2}$$

and the three positive numbers α_0, α_1 and β_1 are the parameters of the process. The variable ε_i is identically and independently distributed (i.i.d.) and in general drawn from the normal distribution $N(0,1)$. The normality of ε is not essential for the discussion below, the only needed property is that the scale for ε_i is fixed by $E[\varepsilon_i^2] = 1$. The integer index i represents a time given by $t = t_0 + i\,\delta t$, where t_0 is an arbitrary starting point, and δt is the process increment. This implicit δt fixes a time scale for the process and for the parameters. The process is assumed to be stationary, and with finite variance $E[\sigma_i^2] < \infty$, which implies $\alpha_1 + \beta_1 < 1$.

With the above parametrization, the properties of the process are entangled in the three parameters. In order to make the properties of the process appear directly in the equation, we rewrite the process for σ_i (equation 2) in the form

$$\begin{aligned} \sigma_i^2 &= \sigma^2(1-\mu_{\mathrm{corr}}) + \mu_{\mathrm{corr}}\left(\mu_{\mathrm{ema}}\sigma_{i-1}^2 + (1-\mu_{\mathrm{ema}})r_{i-1}^2\right) \\ &= \sigma^2 + \mu_{\mathrm{corr}}\left(\mu_{\mathrm{ema}}\sigma_{i-1}^2 + (1-\mu_{\mathrm{ema}})r_{i-1}^2 - \sigma^2\right) \end{aligned} \tag{3}$$

with the change of parameters

$$\sigma^2 = \frac{\alpha_0}{1-\alpha_1-\beta_1} \tag{4}$$

$$\mu_{\mathrm{corr}} = \alpha_1+\beta_1 \tag{5}$$

$$\mu_{\mathrm{ema}} = \frac{\beta_1}{\alpha_1+\beta_1} \tag{6}$$

This reparametrization exemplifies the properties of the process, in particular

$$E\left[\sigma_i^2\right] = E\left[r_i^2\right] = \sigma^2 \tag{7}$$

$$E\left[\sigma_{i+k}^2 \mid \sigma_i^2\right] = \sigma^2 + \mu_{\mathrm{corr}}^k\left(\sigma_i^2 - \sigma^2\right) \tag{8}$$

where σ^2 is the mean of r_i^2 and σ_i^2. The parameter μ_{corr} corresponds to the exponential decay of the conditional mean volatility. The parameter μ_{ema} is acting somewhat similarly to an exponential moving average (ema). The decay of the correlation for the return and latent volatility are also given by μ_{corr}

$$\begin{aligned} \rho[\sigma^2](k) &= \frac{E[\sigma_{i+k}^2\,\sigma_i^2] - E[\sigma_i^2]^2}{E[\sigma_i^4] - E[\sigma_i^2]^2} = \mu_{\mathrm{corr}}^k \\ \rho[r^2](k) &= \frac{E[r_{i+k}^2\,r_i^2] - E[r_i^2]^2}{E[r_i^4] - E[r_i^2]^2} \\ &= \mu_{\mathrm{corr}}^k \frac{\kappa_4\left(\mu_{\mathrm{ema}} + (1-\mu_{\mathrm{ema}})E[\varepsilon^4]\right) - 1}{\kappa_4 E[\varepsilon^4] - 1} \end{aligned} \tag{9}$$

with

$$\kappa_4 = \frac{E[\sigma^4]}{\sigma^4} = \frac{1-\mu_{\mathrm{corr}}^2}{1-\mu_{\mathrm{corr}}^2\left(1+(1-\mu_{\mathrm{ema}})^2(E[\varepsilon^4]-1)\right)}\,. \tag{10}$$

Throughout this paper, the volatility is annualized by using

$$\sigma_{\mathrm{ann}}^2 = \frac{1\mathrm{year}}{\delta t}\sigma^2\,. \tag{11}$$

This makes σ_{ann} independent of δt. The same transformation is used for the latent volatility σ_i and for the return r_i. In this paper, the computations are done with daily data $\delta t = 1$ day, with 1 year = 250 days, and the time intervals are expressed in days.

Each parameter μ acts multiplicatively and, therefore, defines an exponential time scale. This is made explicit by a further change of coordinates

$$\begin{aligned} \mu_{\mathrm{corr}} &= \exp\left(-\delta t/\tau_{\mathrm{corr}}\right) \\ \mu_{\mathrm{ema}} &= \exp\left(-\delta t/\tau_{\mathrm{ema}}\right) \end{aligned} \tag{12}$$

where τ_{corr} and τ_{ema} correspond to the exponential decay time interval of the correlation and ema. As often, time scales vary widely, it is therefore useful to introduce other coordinates as the logarithm of the time intervals

$$\begin{aligned} z_{\text{corr}} &= \ln(\tau_{\text{corr}}) \\ z_{\text{ema}} &= \ln(\tau_{\text{ema}}) \end{aligned} \tag{13}$$

such that the z parameterization corresponds to a double logarithmic transformation of the μ parameters, i.e. $\mu = \exp(-\delta t \exp(-z))$.

These successive transformations define four coordinate systems on the parameter space, namely $(\alpha_0, \alpha_1, \beta_1)$, $(\sigma_{\text{ann}}, \mu_{\text{corr}}, \mu_{\text{ema}})$, $(\sigma_{\text{ann}}, \tau_{\text{corr}}, \tau_{\text{ema}})$ and $(\sigma_{\text{ann}}, z_{\text{corr}}, z_{\text{ema}})$. These different coordinates emphasize that the GARCH(1,1) process is parametrized by a three dimensional space, on which we can use different coordinate systems. Generically, we use θ to denote one point in the parameter space, and practically θ is represented in a given coordinates system.

The domains for the coordinates are

$$\begin{aligned} 0 &< \alpha_0, \alpha_1, \beta_1 \qquad \alpha_1 + \beta_1 < 1 \\ 0 &< \sigma \\ 0 &< \mu_{\text{corr}}, \mu_{\text{ema}} < 1 \\ 0 &< \tau_{\text{corr}}, \tau_{\text{ema}} \end{aligned} \tag{14}$$

and no constraints on $z_{\text{corr}}, z_{\text{ema}}$. One advantage of these new coordinates is that the domains are becoming simpler in the new coordinate systems. Yet, as a guard for the numerical search of the maximum likelihood, sufficiently large finite bounds may still be imposed to prevent overflow or zero values, particularly because the coordinate changes involve exponentials and logarithms.

3 The log-likelihood

Usually, the parameters θ are fitted by a (pseudo) log-likelihood procedure. Assuming that the random variables ε_i have a Gaussian distribution, the log-likelihood function is

$$l(\theta) = -\frac{1}{2n} \sum_i \left(\ln(2\pi) + \ln(\sigma_i^2) + \frac{r_i^2}{\sigma_i^2} \right), \tag{15}$$

where n is the number of data points in the sample. An initial fraction of data must be used for build-up, because of the memory of the volatility process. An estimate $\tilde{\theta}$ for the parameters is given by the solution of the maximization problem

$$\max_{\theta} \ l(\theta). \tag{16}$$

Many desirable properties are known when fitting parameters by a log-likelihood procedure[2]. The most generally applicable property is the independence on the coordinate system: the estimation can be done in any parametrization and the results will be identical, up to the reparametrization. This property is true for finite samples and any data set (assuming a non-degenerate maximum). In particular, if the process is misspecified (i.e. the data were not generated by the fitted process), the maximum is identical in any coordinate system.

[2] A general reference on this topic is (Davidson and MacKinnon, 1993).

The convergence properties of the algorithm used to solve the max equation change depending on the coordinate system. Essentially, all efficient maximization algorithms use a Taylor expansion around the maximum. As a heuristic estimate, the convergence is faster for a larger domain in which the second order Taylor expansion around the solution is a good approximation. Practically, we use in this paper a BFGS algorithm, according to the Numerical Recipes (Press et al., 1992). For this section, the empirical data set used to compute the figures is the daily foreign exchange (FX) rate for USD/CHF. The year 1994 is used for build-up[3], and the years 1995 to 1998 for the estimations. We have checked that the properties discussed in this paper are robust against changes of the data set (other currency pairs, equities indices, or bond indices), or changes of the time period, or of the build-up and sample lengths.

The properties of the $l(\theta)$ function are illustrated by plotting its values along cuts in the parameter space. Those cuts are done along constant coordinate planes, and this is where the different coordinate systems play an important role. A cut in the $(\sigma, \mu_{\text{corr}})$ plane at constant μ_{ema} is presented in Figure 1. This figure does not really look like a nice parabolic maximum. Let us emphasize that the level lines for the contour plot are not regularly spaced, but much closer around the maximum. In fact, the maximum is very flat along an ‘L’ shape. A closer inspection shows that something happens along the lower side of the figure, but the log-likelihood changes too rapidly in these coordinates to display a clear figure.

The same cut, with the same domain but in the $(\sigma, \tau_{\text{corr}})$ plane, is presented on Figure 2. Here, the structure starts to unravel: the large flat area that occupies most of the figure is the structure that was wedged along the $\mu_{\text{corr}} = 1$ axis. The maximum is now at the very top of the figure, at $\tau_{\text{corr}} \simeq 20$ days, and is still very asymmetric.

In order to see the maximum clearly, another logarithmic transformation is necessary. The same cut, but in the $(\sigma, z_{\text{corr}})$ plane, is presented on Figure 3. Here, the structures completely unravel: *a long ridge of nearly optimal solutions exists.* This one dimensional quasi-degeneracy is the major cause of trouble for the maximization algorithm. This feature is not visible in the (α, β) coordinates because the ridge is nearly perpendicular to the μ_{ema}, τ_{ema} or z_{ema} coordinates. An arbitrary cut, in particular in the (α, β) coordinates, shows only an isolated maximum. The comparison of the three figures also explains why, generically, the maximization algorithm performs much better when using the $(\sigma_{\text{ann}}, z_{\text{corr}}, z_{\text{ema}})$ coordinates: the log-likelihood is much better approximated by its second order Taylor expansion.

In order to get a better quantitative picture for these quasi degenerate solutions, we solve for a given σ the maximization problem for $(z_{\text{corr}}, z_{\text{ema}})$. Then, we can plot the solution $(\tilde{z}_{\text{corr}}, \tilde{z}_{\text{ema}})$ as a function of σ, as shown on Figure 4. The long ridge is now clearly visible as a plateau on the log-likelihood function. Note that the relative difference in the log-likelihood between the best solution and the plateau is only of 0.3%. By comparison, computing with the same time series (USD/CHF, with 1 year for build up and 4 years for the fits), but sliding the starting point for the fit every month from 1986 to 1993, leads to a variation of the log-likelihood of 10%.

The horizontal plateau corresponds to $|\partial l(\theta)/\partial\theta_k| \simeq 0$, and possibly fools an optimization algorithm to fulfill its convergence criterion. This can be a major source of spurious solutions. Moreover, a large amount of data was used in order to compute this figure, yet using less data can create secondary maxima or possibly another absolute maximum. These spurious solutions are worrisome as the corresponding parameters are completely meaningless. For example, the average volatility of the resulting process (which is given directly by σ) can be too large by a factor 10.

[3] The GARCH processes have an internal state with memory, namely σ_i. The internal state is initialized with an estimate of the mean volatility, and then one year of data is used to build-up the internal state. In this way, at the begining of the sample used to estimate the log-likelihood function, σ_i has an exponentialy small dependency on the initial condition.

PSfrag replacements

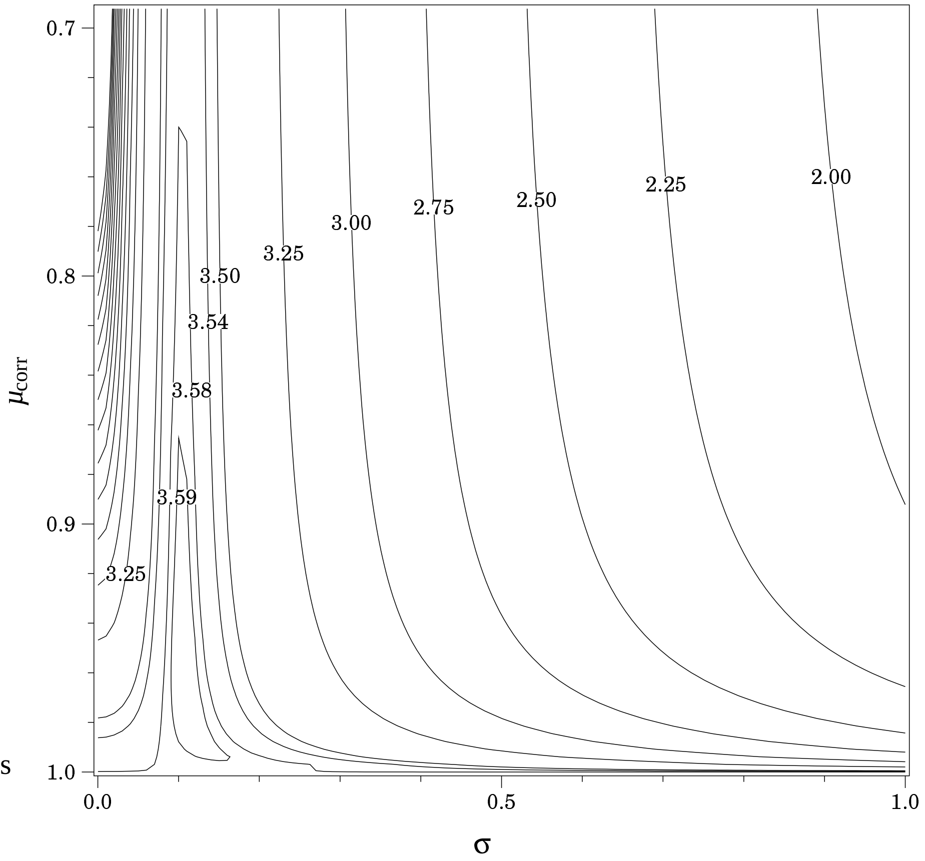

Figure 1: The log-likelihood as a function of σ_{ann} and μ_{corr}, for $\mu_{\mathrm{ema}} = 0.9514 = \exp(-\exp(-3))$. The data are USD/CHF foreign exchange, with 1 year of build-up for the process, and 4 years for the computation of the log-likelihood. Note that the levels for the contour plot are much narrower close to the maximum.

PSfrag replacements

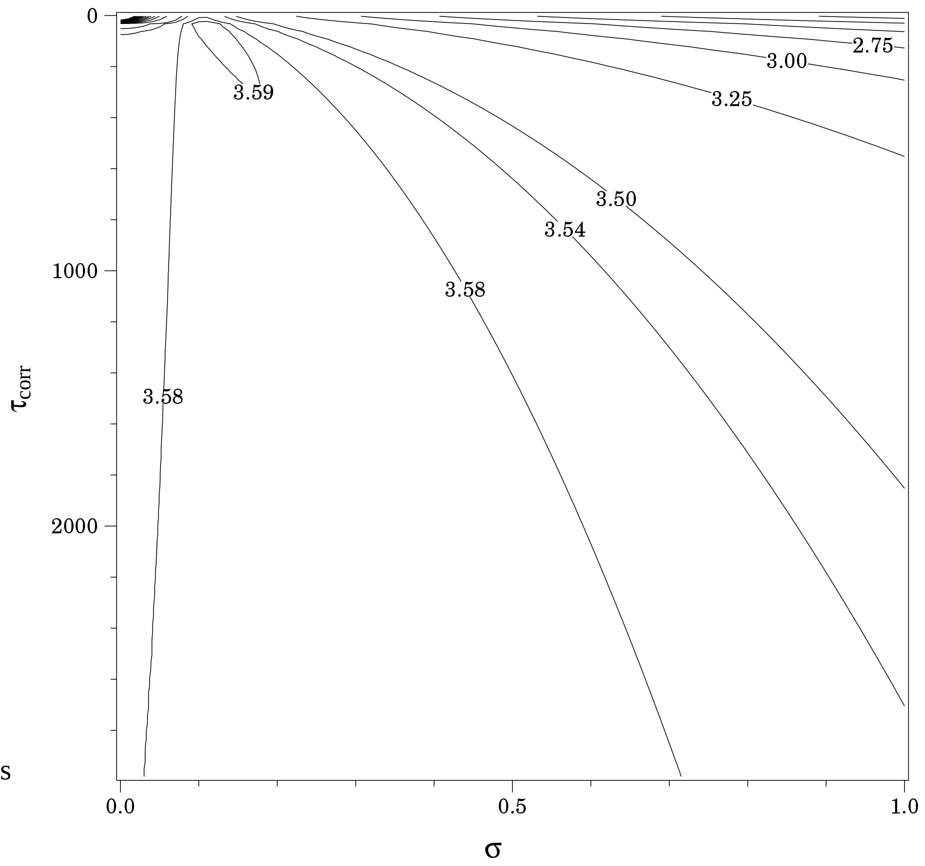

Figure 2: The log-likelihood as a function of σ_{ann} and τ_{corr}, for $\tau_{\text{ema}} = 20.08 = \exp(3)$. The data, domain for the parameters, and levels for the contour plot are as in Figure 1.

PSfrag replacements

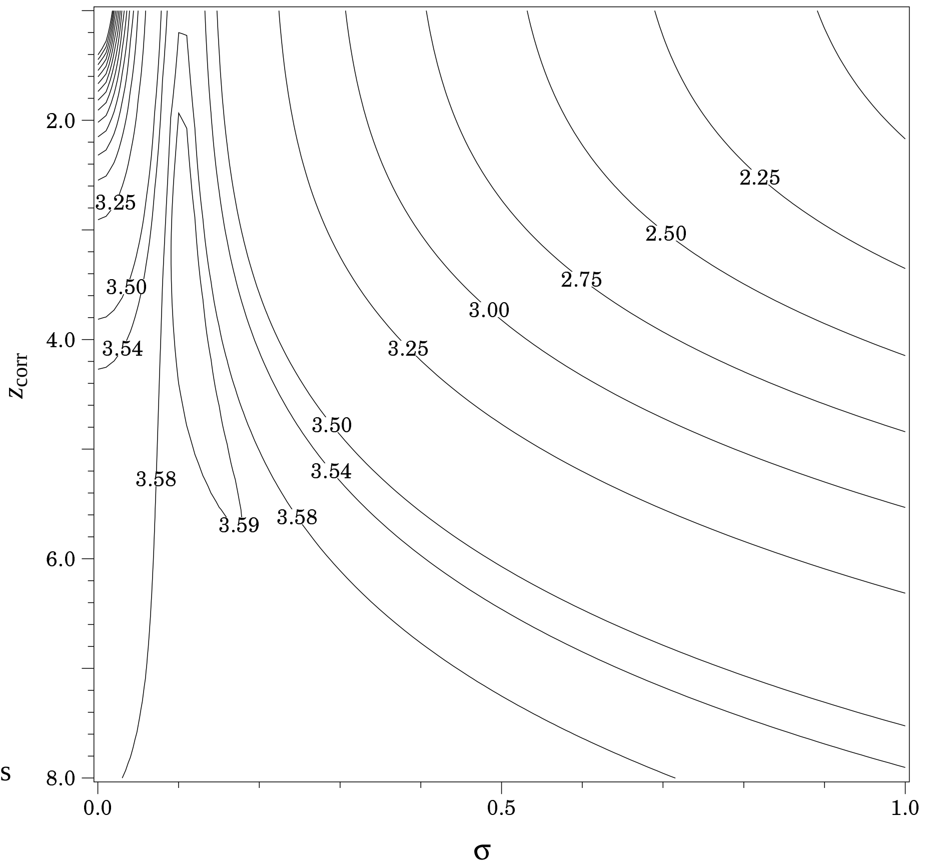

Figure 3: The log-likelihood as a function of σ_{ann} and z_{corr}, for $z_{\text{ema}} = 3$. The data, domain for the parameters, and levels for the contour plot are as in Figure 1.

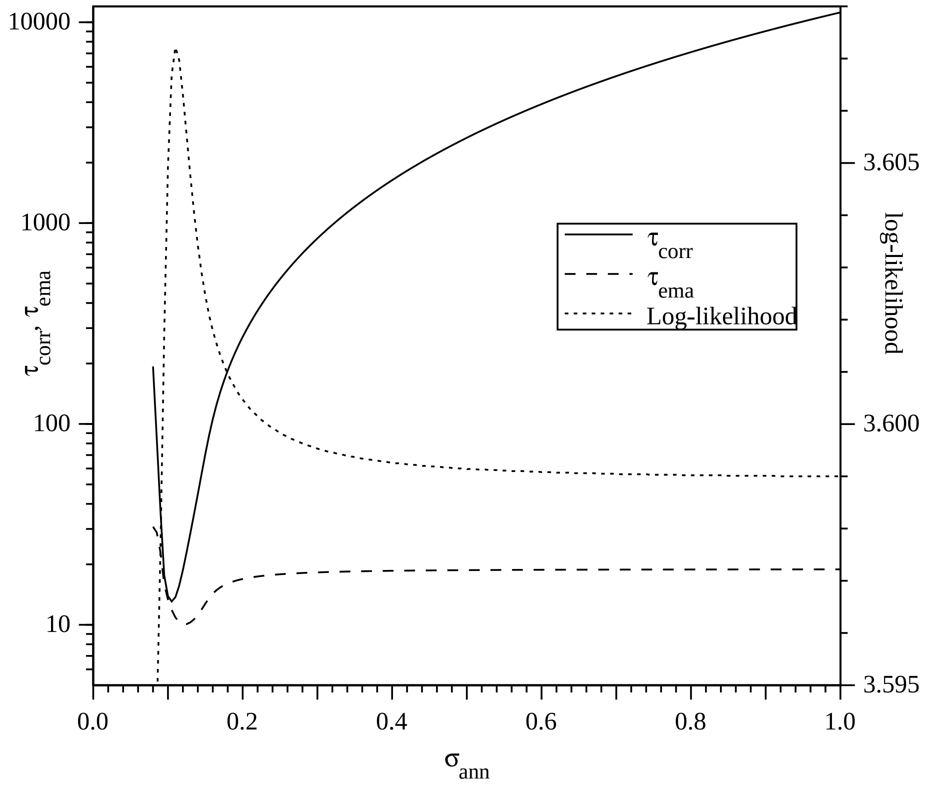

Figure 4: $\tilde{\tau}_{\text{corr}}$, $\tilde{\tau}_{\text{ema}}$ and the log-likelihood as a function of σ_{ann}. The data are as in Figure 1.

These spurious solutions are not easy to detect in the usual (α, β) coordinates, their most visible signature being $\tilde{\alpha}_1 + \tilde{\beta}_1 \simeq 1$.

In order to define a more robust fitting procedure for the parameters, we propose to use a mixed estimation, with a moment estimate for σ and with a log-likelihood for $z_{\text{corr}}, z_{\text{ema}}$. More precisely, σ is computed from the moment estimate

$$\tilde{\sigma}^2 = \langle r_i^2 \rangle \tag{17}$$

where $\langle \cdot \rangle$ denotes the sample average. The remaining parameters are obtained by solving the restricted maximum likelihood problem

$$\min_{(z_{\text{corr}}, z_{\text{ema}})} l(\tilde{\sigma}, z_{\text{corr}}, z_{\text{ema}}) \ . \tag{18}$$

In order to differentiate from the usual ‘log-likelihood fit all’ procedure, we call the resulting two steps estimate a *restricted* estimate, because for the log-likelihood step, the parameter space of the process is restricted to the subspace corresponding to the observed volatility. This method enforces a natural soundness criterion on the fitted process, namely the volatility of the fitted process is identical to the volatility of the data. Moreover, this is important for long term forecasts (see equation 8) made with this process. Note that some software packages like S+ already allow the estimation of a volatility like constraint with a moment estimator. Yet, this is used with the original (α, β) parametrization to add a constraint that essentially fixes α_0. Therefore, there is no change of coordinates involved in this procedure.

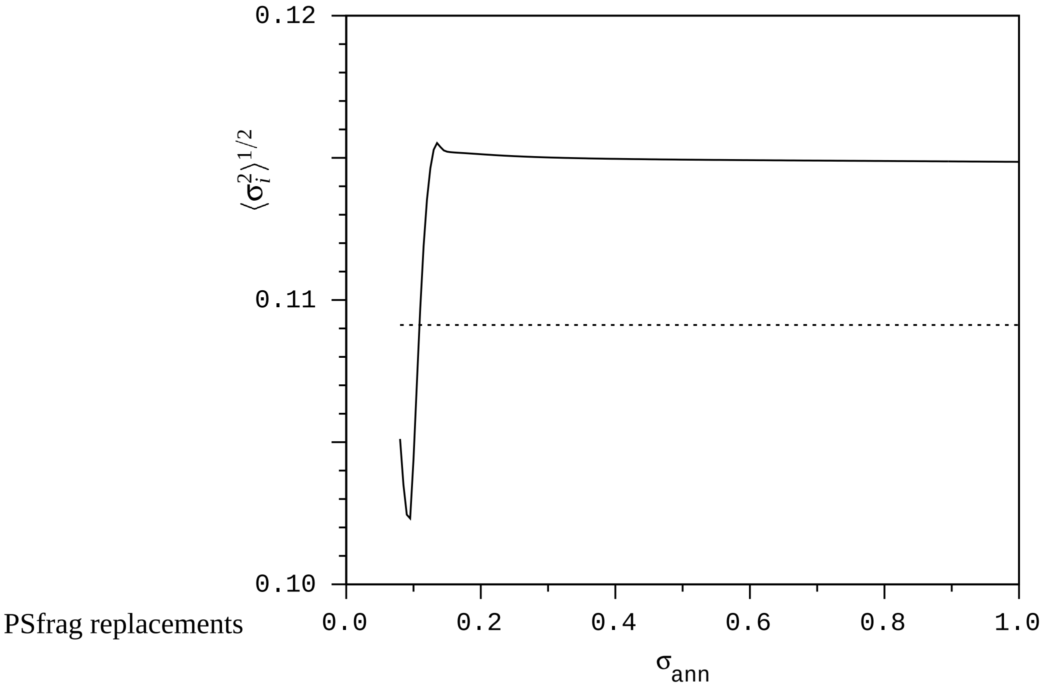

Figure 5: The annualized latent volatility $\langle\sigma_i^2\rangle^{1/2}$ (full line) and annualized data volatility $\langle r_i^2\rangle^{1/2}$ (dotted line) for the GARCH(1,1) process with different values for the parameter σ_{ann}. At the point where the two curves intersect, the process parameter σ_{ann} is such that the latent volatility is equal to the data volatility. The data are as in Figure 1.

4 The causes of the trouble

We still need to understand why we obtain this family of almost degenerate solutions. The important point to realize is that, when fitting a process on data, the three quantities σ^2, $\langle\sigma_i^2\rangle$ and $\langle r_i^2\rangle$ are different, namely the identity in equation 7 does not hold anymore for *sample average*. The process for σ_i is still defined by equation 3, and we can take its unconditional average. After straightforward calculations, the unconditional average can be written

$$\sigma^2 = \langle r_i^2\rangle + \left(\frac{\mu_{\mathrm{corr}}(1-\mu_{\mathrm{ema}})}{1-\mu_{\mathrm{corr}}} + 1\right)\left(\langle\sigma_i^2\rangle - \langle r_i^2\rangle\right) \tag{19}$$

The quantity $\langle r_i^2\rangle$ is a number depending only on the data, the quantity $\langle\sigma_i^2\rangle$ is still a function of the process parameters. In the above equation, if $\langle\sigma_i^2\rangle = \langle r_i^2\rangle$, then the usual identity between all variances is true. Similarly to Figure 4, we plotted on Figure 5 the quantities $\langle\sigma_i^2\rangle$ and $\langle r_i^2\rangle$ along the almost degenerate solutions, using σ_{ann} as a parameter. On this figure, the difference between the three variances is clear. Moreover, $\langle\sigma_i^2\rangle$ is almost constant along the branch of near solutions, but different by about 5% from $\langle r_i^2\rangle$. Only at the maximum likelihood solution, the equality $\langle\sigma_i^2\rangle = \langle r_i^2\rangle = \tilde{\sigma}^2$ is fulfilled to a high degree of accuracy. Therefore, the second term in equation 19 is non zero, and the $1/(1-\mu_{\mathrm{corr}})$ creates a singularity for σ. In order to check for this explanation, we approximate $\langle\sigma_i^2\rangle$ by a constant and plug the appropriate numbers in equation 19 in order to obtain an estimate for σ. Then, we compare σ versus μ_{corr} as obtained along the nearly degenerate solution, and σ versus μ_{corr} as obtained from equation 19: the two curves closely match, in agreement with the above analysis.

An interesting question is to determine whether this one dimensional degeneracy is originating in the GARCH process itself, or in the inability of the GARCH process to fit financial data (say for example, because the GARCH(1,1) process has an exponential correlation decay whereas financial data have a power law decay). To settle this question, we generated synthetic data with a GARCH(1,1) process. Then, we plotted as above the log-likelihood landscape of a GARCH(1,1) process, but using return from the synthetic GARCH(1,1) data. A similar picture emerges, indicating that this one dimensional degeneracy is a property of the GARCH process.

We fitted a GARCH(1,1) process using a restricted estimate on more than 200 daily time series, including foreign exchange rates, interest rates, bond indices and equities indices. The annualized volatility changes significantly from 1% to 50%, particularly between families of assets. Typical time decay parameters are $1 \leq z_{\mathrm{corr}}, z_{\mathrm{ema}} \leq 4$, or 2.7 days $\leq \tau_{\mathrm{corr}}, \tau_{\mathrm{ema}} \leq$ 54 days. Parameters which are very different from those ranges should be considered *a priori* as suspicious. We also fitted higher frequency data for USD/CHF on the GARCH(1,1) process (see section 6). The fit is increasingly dubious with higher frequency. Yet, the restricted fit gives better estimated parameters, with a log-likelihood differing by less than 0.0025%.

5 Estimation on a finite sample

When introducing the log-likelihood procedure, we mentioned that the fit is independent of the coordinate system, even for finite samples and misspecified processes. More desirable properties of the log-likelihood procedure are true when making more assumptions. First, let us assume that the data generating process is given, and that we are fitting these data on the same process. This hypothesis enforces consistency, namely that we are fitting the right data generating process. Then, it is known that asymptotically (i.e. when the sample size goes to infinity), the fitted parameters converge in probability to the true parameters and have a Gaussian distribution around the true values. Moreover, the standard deviation for the Gaussian probability density function (pdf) is related to the information matrix, and decays as $1/\sqrt{n}$ where n is the sample size (Davidson and MacKinnon, 1993). Under a change of coordinates, the information matrix changes in the natural way, namely by conjugation with the Jacobian matrix of the coordinate change. Therefore, a change of coordinates does not modify the asymptotic convergence properties of the log-likelihood fit. Another important property of log-likelihood fit is its efficiency, namely the estimator is in some sense optimal (Davidson and MacKinnon, 1993), and this is true in any coordinate system.

Yet, beside the invariance with reparametrization, very little is known about log-likelihood estimates using finite data samples. This is of important practical concern because, typically, a few years of daily data are fitted on a process, meaning a few hundred to a few thousand points. In particular, the finite sample estimates are generically biased. Moreover, under nonlinear transformations, at most one coordinate system can be unbiased (because $\langle f(x) \rangle \neq f(\langle x \rangle)$ for a non linear function f). Therefore, the various coordinate systems introduced previously for the GARCH(1,1) process have different biases.

In this section, we study the finite size distribution of the fitted parameters, and the bias of the various coordinates. For this purpose, we generate samples of data of given length n with a GARCH(1,1) process with parameters corresponding to $\sigma_{\mathrm{ann}} = 10\%$, $z_{\mathrm{corr}} = 3$ and $z_{\mathrm{ema}} = 2.5$, or for the other coordinates $\alpha_0 = 1.94310^{-6}$, $\alpha_1 = 0.0750$, $\beta_1 = 0.8764$, $\mu_{\mathrm{corr}} = 0.9514$, $\mu_{\mathrm{ema}} = 0.9212$, $\tau_{\mathrm{corr}} = 20.2$, $\tau_{\mathrm{ema}} = 12.18$. These values correspond to typical FX parameters. Generically, a coordinate of the data generating process is denoted by θ_0. Then, a GARCH(1,1) process is fitted on this data set, resulting in an estimate θ for each coordinate. This procedure is repeated $N = 100{,}000$ times, and the means $\overline{\theta}$, the standard deviation[4] $\mathrm{stdDev}(\theta)$, and the empirical pdf for

[4] We denote the standard deviation by stdDev to avoid confusion with the process parameter σ.

the fitted parameters θ is computed. This procedure is repeated for various sample sizes n.

The relative bias for the various coordinates is estimated as follows: for a coordinate θ, we make the ‘Ansatz’

$$\overline{\theta}(n) = \theta_0 \left(1 + \frac{b_\theta}{n}\right) \tag{20}$$

where $b_\theta = b_\theta(n)$ is the relative bias and n the sample size. This form is dictated by the asymptotic (unbiased) Gaussian distribution for the rescaled parameters, which implies that the bias has to decay asymptotically faster than $1/\sqrt{n}$. In order to check that this form captures the leading n dependency, we display on Figure 6 the relative bias for the sample sizes in the range $125 \leq n \leq 2000$, which corresponds to a range of 6 months to 8 years of daily data. Some of the coordinates $(\beta_1, \sigma_{\mathrm{ann}}, \mu_{\mathrm{corr}}, \mu_{\mathrm{ema}}, z_{\mathrm{corr}})$ seem to be already close to the asymptotic behavior. The other coordinates have still a clear $1/\sqrt{n}$ correction. A study of the standard deviations and empirical probability density functions of the scaled deviation $\delta\theta = \sqrt{n}\,(\theta - \theta_0)$ points to similar finite size behaviors. For example, the pdf for $\delta\alpha_0$ is clearly skewed, even for a sample size of $n = 2000$, which can be thought as large enough. This slow convergence to the asymptotic behavior originates in the correlation of the process, as measured by τ_{corr}. Roughly, an independant estimate for the parameter is obtained after τ_{corr} data, reducing the sample size n to an effective sample size $n_{\mathrm{eff}} = n/\tau_{\mathrm{corr}}$. For example, one year of daily data corresponds to $n_{\mathrm{eff}} \simeq 250/20 \simeq 12$, which is clearly a very small sample. Practically, we are very often fitting processes in this small effective sample size and therefore we should expect strong corrections to the asymptotic theoretical results. This becomes particularly relevant for inference and hypothesis testing, and in this respect, the choice of coordinates *does* matter. In the context of the Wald test (Davidson and MacKinnon, 1993), the dependency of the test with respect to the algebraic formulation of the null hypothesis has been studied by (Phillips and Park, 1988). For example, in view of the Figures 1, 2 and 3, this issue seems important when testing for the unit root $\alpha_1 + \beta_1 = \mu_{\mathrm{corr}} = 1$. However, hypothesis testing would bring us beyond the scope of the present paper. Returning to the Figure 6, we see that α_0, τ_{corr} and τ_{ema} are strongly biased, α_1, β_1, μ_{corr} and z_{corr} have a medium bias, and σ_{ann}, μ_{ema} and z_{ema} have a small bias. This is yet another reason not to use the usual (α_0, α_1, β_1) coordinates, but to prefer the $(\sigma, \mu_{\mathrm{corr}}, \mu_{\mathrm{ema}})$ or $(\sigma, z_{\mathrm{corr}}, z_{\mathrm{ema}})$ coordinates.

With an estimate for the standard deviation of $\overline{\theta}$

$$\mathrm{stdDev}(\overline{\theta}) = \sqrt{\frac{\overline{\theta^2} - \overline{\theta}^2}{N}}\,, \tag{21}$$

the standard deviation of the relative bias can be estimated $\mathrm{stdDev}(b) = n\,\mathrm{stdDev}(\overline{\theta})/\theta_0$. Then, at a 10 sigma level, no coordinate is free of biases. Another comment about fitting small samples must be added. In order to solve the maximization problem, our best fitting algorithm is used, namely a BFGS algorithm working in the $(\sigma_{\mathrm{ann}}, z_{\mathrm{corr}}, z_{\mathrm{ema}})$ coordinates. The search algorithm is started at the data generating values, with possible restart at 6 neighboring points when a solution is not found. Despite this fairly robust maximization algorithm, a fraction of the data set *never* converges to a meaningful solution, but runs to ‘infinity’ (i.e. to the boundaries of the domains set to avoid over- and underflow). For the data sets of size 125, 250, 500, 1000 and 2000, the fractions of unconverged solutions are respectively 24%, 7%, 1%, 0.03% and 0%. This indicates that, for these fractions of the data sets, there is no global maximum (despite the fact that we are fitting the data generating process). This is yet another warning about potential problems of estimates made on small data sets.

PSfrag replacements

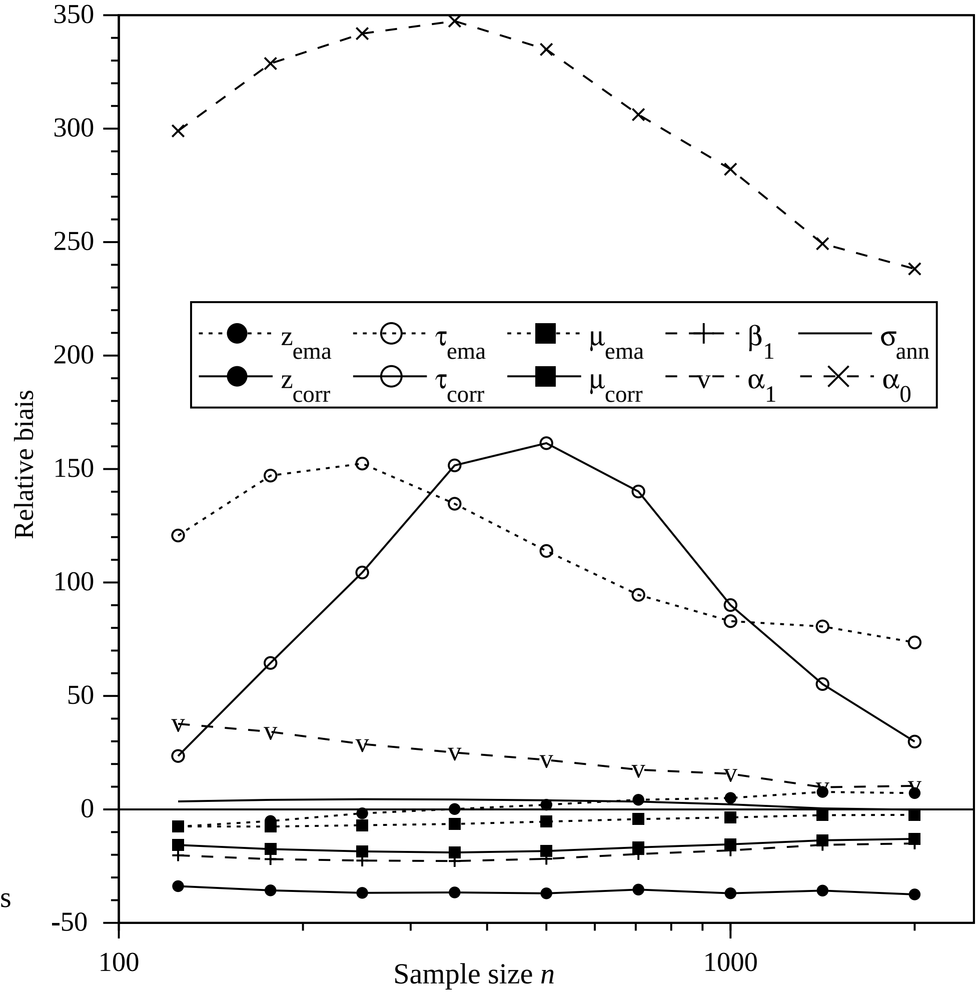

Figure 6: The relative bias versus the sample size, for the various coordinates.

6 Estimation at higher frequencies

Up to this point, the fitting of daily data is discussed. In this section, the above issue is explored at higher frequencies. A first study using intra-day data can be found in (Guillaume et al., 1994). In order to discount the daily and weekly seasonalities, a synthetic homogeneous time series is computed, sampled regularly in dynamical business time scale (Breymann et al., 2000). This time scale measures the recent intra-week volatility pattern and builds a time scale flowing at a pace related to this pattern, similarly to the theta-time proposed in (Dacorogna et al., 1996). Moreover, Holidays and Daylight Saving Time are taken into account. The regular time series is computed from tick-by-tick quotes for USD/CHF and covers 11 years, from 1.1.1989 to 1.1.2000. The synthetic time series is sampled every 14 minutes in dynamical business time, with a linear interpolation between ticks, and with the logarithmic middle price as the value. The year 1989 is used for build-up of the GARCH(1,1) process and the decade 1990 to 2000 for the fit.

With this regular time series, it is easy to compute the return at different frequencies, and to investigate the fitted GARCH(1,1) process. As a by-product, we can compare the fitted parameters with the aggregation relation for the GARCH(1,1) process as derived by (Drost and Nijman, 1993). In this paper, assuming a GARCH(1,1) process at scale δt, the authors compute the parameters for the equivalent GARCH(1,1) process at scale $m\delta t$. These relations are quite complicated, yet they simplify drastically for σ_{ann} and τ_{corr}

$$\begin{aligned} \sigma_{\text{ann}}(m\delta t) &= \sigma_{\text{ann}}(\delta t) \\ \tau_{\text{corr}}(m\delta t) &= \tau_{\text{corr}}(\delta t). \end{aligned} \tag{22}$$

In this form, both relationships are very easy to check. The decimation equation for the last parameter seems not to simplify (it involves solving an implicit equation). At the highest frequency, we draw the equivalent of Figures 1, 2 and 3. Essentially, the same problem is present, but the manifold of nearly degenerate solutions is much flatter. This makes the fit of high frequencies data even more difficult.

The GARCH(1,1) processes is estimated on the above data, at time intervals ranging from 14 minutes to 8 days. The fit is done with a log-likelihood computed with a Student-t probability density for the residuals, and the number of degrees of freedom is also optimized. The resulting volatilities are displayed in Figure 7. The agreement between the 3 volatilities is excelent above $\sim$1 day, but deteriorates with smaller time intervals. Figure 7 indicates that at the maximum likelihood, the equality $\sigma^2 = \langle r_i^2 \rangle$ is not fulfilled for intra-day time intervals. One way to interpret this result is that GARCH(1,1) is not able to describe such high frequency data.

In order to enforce the identity $\sigma^2 = E[r_i^2]$, the restricted fitting procedure is used. As the annualized data volatility $E[r[\delta t]^2]$ is roughly constant, this enforces that $\sigma_{\text{ann}}(\delta t)$ is also almost constant. At the maximum likelihood, the relative difference of the log-likelihood between restricted and unrestricted estimates is in the worst case 0.0025% (at the highest frequency). This illustrates once more the difficulty of such high frequency fits and the flatness of the log-likelihood in the parameter space. The resulting τ parameters are displayed in Figure 8. For time intervals longer than $\sim$0.7 day, the decay correlation time τ_{corr} is approximately constant, with a value $\sim$ 90 days. This behavior is consistent with the aggregation properties of GARCH(1,1), and points to the existence of a preferred correlation decay frequency, of the order of 3 months. Yet, the situation changes for time intervals smaller than 0.7 day. In this domain, the correlation time is very well described by a linear dependency with the return time interval $\tau_{\text{corr}} \sim n\delta t$, in contradiction with the theoretical aggregation properties. Let us note that, during the working days, 0.7 day in a business time scale corresponds on average to 0.7×5/7 = 0.5 day, namely the inflexion point is around 12h. This change of behavior for the parameters may be explained by several arguments, for example by the

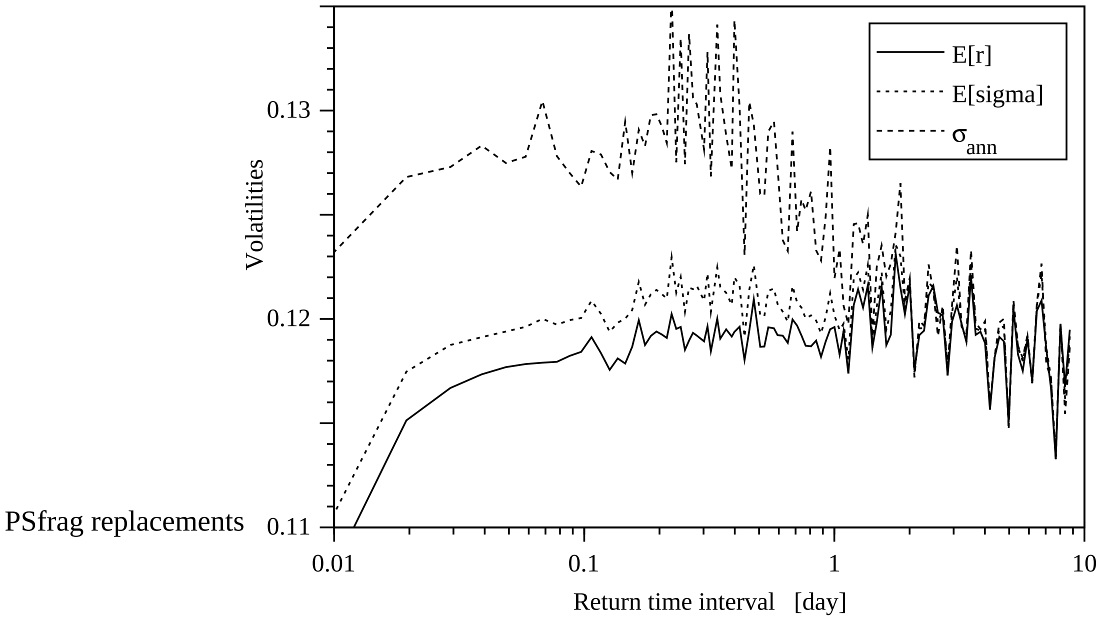

Figure 7: The annualized volatilities $\langle r_t^2\rangle^{1/2}$, $\langle\sigma_t^2\rangle^{1/2}$ and σ_{ann} versus the return time interval δt.

presence of intra-day speculators, or by the inadequacy of the GARCH(1,1) process to describe intra-day data.

It is also worthwhile to analyse the number of degrees of freedom for the Student-t distribution, as reported on Figure 9. For the whole range of return time intervals, the values are between 4.5 to 9, with a value ∼6 for $\delta t = 1$ day. These values are too small to consider a Gaussian distribution as a good approximation of the Student distribution. Besides, there is no apparent change of regime between deep intra-day return, and daily return. This points to the fact that a Student distribution should be used at all frequencies, including for daily data.

7 Generalization for GARCH(p,q)

The generalization of the coordinates change to GARCH(p,q) is fairly straighforward, at least for the transformation to the $(\sigma, \mu_{\text{corr}}, \mu_{\text{ema}})$ coordinates. The next coordinate change for τ requires more sophistication. In order to simplify the notation, we consider the model GARCH(m,m) with $m = \max(p,q)$, and possibly some of the coefficients are set to zero:

$$\sigma_i^2 = \alpha_0 + \sum_{k=1}^{m} \alpha_k r_{i-k}^2 + \sum_{k=1}^{m} \beta_k \sigma_{i-k}^2 \tag{23}$$

The equations 4 to 6 become

$$\begin{aligned} \sigma^2 &= \frac{\alpha_0}{1-\sum_{k=1}^m \alpha_k + \beta_k} & (24)\\ \mu_{\text{corr},k} &= \alpha_k + \beta_k & (25)\\ \mu_{\text{ema},k} &= \frac{\beta_k}{\alpha_k + \beta_k} & (26) \end{aligned}$$

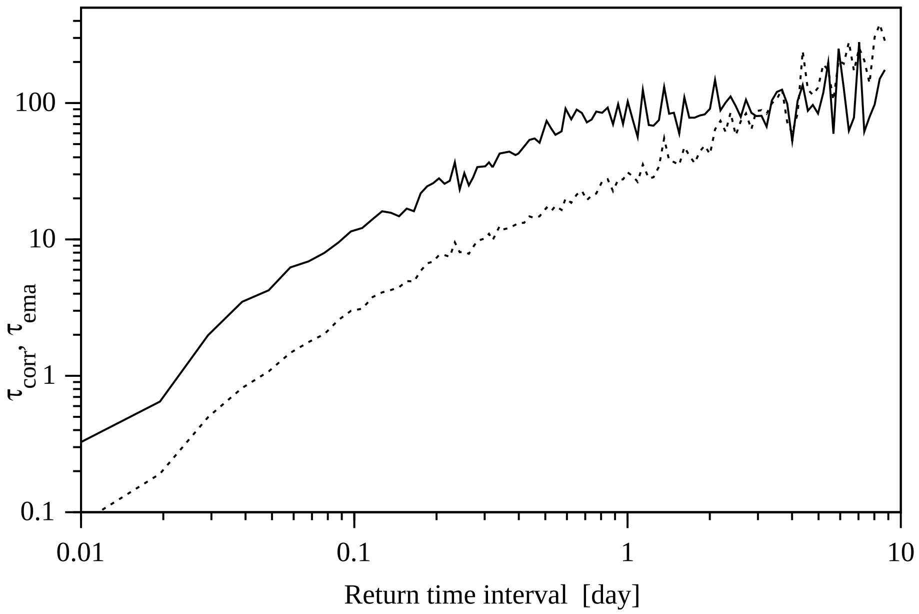

Figure 8: The fitted GARCH(1,1) parameters τ_{corr} (full line) and τ_{ema} (dotted line) versus the return time interval δt.

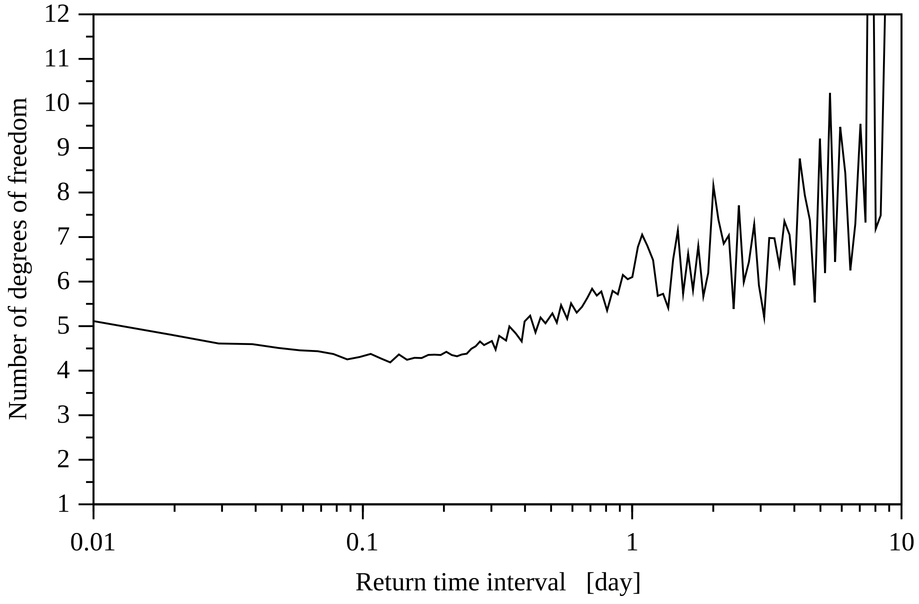

Figure 9: The optimal number of degrees of freedom for the Student-t distribution of the residual, when estimating a GARCH(1,1) process with a (restricted) log-likelihood procedure. The estimate is done for different return time interval δt, given on the x-axis.

and lead to

$$\sigma_i^2 = \sigma^2 + \sum_{k=1}^{m} \mu_{\mathrm{corr},k} \left(\mu_{\mathrm{ema},k}\sigma_{i-k}^2 + (1-\mu_{\mathrm{ema},k})r_{i-k}^2 - \sigma^2\right). \tag{27}$$

A short computation of the unconditional average of the process shows that σ corresponds to the mean volatility $\sigma^2 = E[\sigma_i^2] = E[r_i^2]$, with the condition $\sum_{k=1}^m \mu_{\mathrm{corr},k} < 1$. In order to ensure the positivity of the volatility σ_i^2 in eq. 27, the conditions $\mu_{\mathrm{corr},k} > 0.0$ for all k are also needed. In order to compute the conditional average and lagged correlations, it is convenient to introduce the new variables

$$\begin{aligned} \gamma_i &= \sigma_i^2 - \sigma \\ \delta_i &= r_i^2 - \sigma_i^2 \end{aligned} \tag{28}$$

for which eq. 27 becomes

$$\gamma_i = \sum_{k=1}^{m} \mu_{\mathrm{corr},k} \left(\gamma_{i-k} + (1-\mu_{\mathrm{ema},k})\delta_{i-k}\right). \tag{29}$$

We denote by Ω_i the information set at time i, and by

$$\begin{aligned} \overline{\gamma}_k &= E[\gamma_{i+k}|\Omega_i] \\ \overline{\delta}_k &= E[\delta_{i+k}|\Omega_i] \end{aligned} \tag{30}$$

the conditional expectation at time $i+k$ given the information set Ω_i. For $k > m$, the recursion equation depends only on $\overline{\gamma}$

$$\overline{\gamma}_j = \sum_{k=1}^{m} \mu_{\mathrm{corr},k}\overline{\gamma}_{j-k} \tag{31}$$

Therefore, the asymtotic properties of the process depend only on the coefficients $\mu_{\mathrm{corr},k}$. In order to compute explicitly the time decay of the correlation, we express the last equation in the form of a Markov chain $\vec{\gamma}_j = M\vec{\gamma}_{j-1}$ for the vector $\vec{\gamma}_j = (\overline{\gamma}_j, \overline{\gamma}_{j-1}, \cdots, \overline{\gamma}_{j-m})^T$. with transition matrix

$$M = \begin{pmatrix} \mu_{\mathrm{corr},1} & \mu_{\mathrm{corr},2} & \cdots & \mu_{\mathrm{corr},m} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \end{pmatrix}. \tag{32}$$

The (complex) characteristic time decays of the Markov chain are given by the eigenvalues of the transition matrix. By expanding recursively on the last column the characteristic equation $|M - \lambda \mathbf{1}_m|$ (where $\mathbf{1}_m$ is the m-dimensional unit matrix), the time decays $\tau_{\mathrm{corr},k} = -\delta t / \ln(\lambda_k)$ are related to the root λ_k of the equation

$$-\lambda^m + \sum_{k=1}^{m} \mu_{\mathrm{corr},k}\lambda^{m-k} = 0. \tag{33}$$

For the process to be well behaved, it must decay to the unconditional mean σ, namely all the roots must be inside the unit circle $|\lambda_k| < 1$. These conditions induce restrictions on the space for the parameters, in particular on α_k, β_k. These conditions do not reduce to the one given in

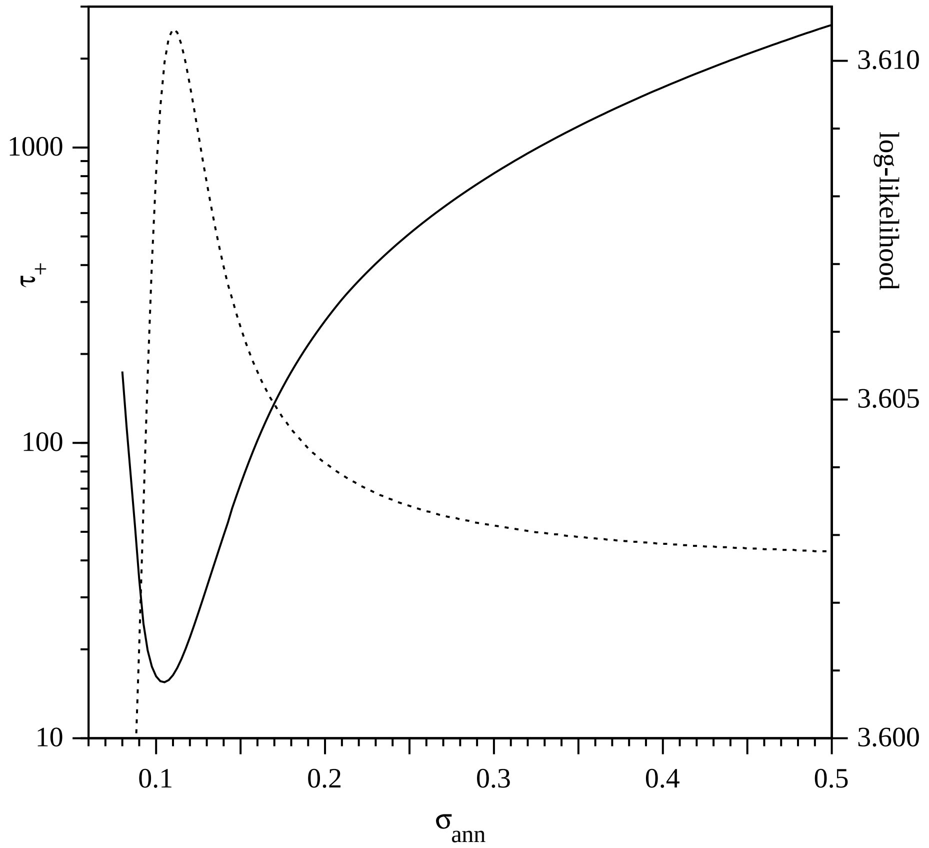

Figure 10: The leading correlation time τ_+ and the log-likelihood as a function of σ_{ann} for the GARCH(2,2) model. The data are as in Figure 1.

(Nelson and Cao, 1992), and as far as we know they are new. Solving the polynomial eq. 33 is clearly the difficult part of the change of coordinates $\mu_{\text{corr},k} \to \tau_{\text{corr},k}$. Then, one more logarithm can be taken to obtain the $z_{\text{corr},k}$ coordinates.

For GARCH(2,2), eq. 33 is quadratic and can be solved by algebraic means. In particular, the above conditions on the $\mu_{\text{corr},k}$ (positivity for all k, sum smaller than one) are sufficient to have a positive determinant, and therefore both roots are real. Moreover, the conditions $|\lambda_k| < 1$ are always satisfied, and therefore do not introduce further restriction.

Intuitively, the manifold of nearly degenerate solution must exist for all GARCH(p,q) models. Similarly to Figure 4 for GARCH(1,1), we have fitted a GARCH(2,2) model on the same data set, for various values of σ_{ann}. The results are displayed in Figure 10 which clearly exhibits the same plateau in the log-likelihood. The sub-optimal solutions correspond to $\sum_{k=1}^{m} \mu_{\text{corr},k} \simeq 1$. For the GARCH(2,2) model, this leads to a leading eigenvalue λ_+ close to one, namely to a very large correlation time τ_+. Therefore the diagnostics for a spurious optimization can similarly be used for GARCH(p,q) models.

A natural question is whether there exist new dangerous directions for a GARCH(2,2) estimate. The most important subspace is spaned by the $(\mu_{\text{corr},1}, \mu_{\text{corr},2})$ coordinate as the correlation properties of the process depend only on these two parameters. A cut in the log-likelihood space through the best solution in the $(\mu_{\text{corr},1}, \mu_{\text{corr},2})$ subspace shows a very elongated parabolic maximum along the boundary $\mu_{\text{corr},1} + \mu_{\text{corr},2} = 1$. The aspect ratio of the parabolic maximum is of order 10 (i.e. the ratio of the curvature). This should not cause trouble for any decent maximization procedure, and therefore there should be no further dangerous direction.

8 Conclusions

The figures clearly indicates the pitfalls present when fitting GARCH(p,q) process. They are related to a one-dimensional manifold of almost degenerate solutions. To avoid possible spurious fits, we suggest:

- using enough data.
- computing σ with the moment estimate $\tilde{\sigma}^2 = \langle r_i^2 \rangle$; then, using a log-likelihood for the remaining parameters.

  maximizing the log-likelihood using the $(\sigma_{\text{ann}}, z_{\text{corr}}, z_{\text{ema}})$ coordinate system, because of the better efficiency of the numerical algorithm. This has the supplementary advantage that there is no constraint on z_{corr} and z_{ema} (except possibly for preventing overflows and underflows).

  checking the results by comparing $\tilde{\sigma}^2$ and $\langle \sigma_i^2 \rangle$.
- If a usual log-likelihood is used instead of the above two step procedure, we suggest checking the validity of the results by comparing $\tilde{\sigma}^2$, $\langle \sigma_i^2 \rangle$ and $\langle r_i^2 \rangle$. Probable spurious solutions correspond to substantial differences between these three quantities, as well as $\sum_k \tilde{\alpha}_k + \tilde{\beta}_k \simeq 1$. For the maximization algorithm, a very small convergence criterion should be taken.

Moreover, we have compared the parameters fitted at various frequencies with the Drost and Nijman (Drost and Nijman, 1993) aggregation relation for GARCH(1,1). Overall, the aggregation relations do not hold, with a weak discrepancy for time intervals longer than 5 hours. This indicates that the GARCH(1,1) does not describe completely financial data.

A direct extension of the above procedure can be done for multivariate processes. In that case, the number of parameters is rapidly growing with the number of time series and the danger of misleading parameter estimations is increasing with the dimension of the parameter space. Estimating volatility-like and correlation-like parameters by a moment estimate seems an efficient way to reduce the complexity and risk of the usual log-likelihood procedure.

9 Acknowledgments

The author would like to thank Tim Bollerslev and Ramazan Gençay for their many suggestions, the members of the research group at Olsen & Associates for their comments, and Ralph Baenziger for his help with the implementation of the optimization algorithms.

References

Bera A. K. and Higgins M. L., 1993, *Arch models: properties, estimation and testing,* Journal of Economic Surveys, **7**(4), 305–362.

Bollerslev T., 1986, *Generalized autoregressive conditional heteroskedasticity,* Journal of Econometrics, **31**, 307–327.

Bollerslev T., Chou R. Y., and Kroner K. F., 1992, *ARCH modeling in finance,* Journal of Econometrics, **52**, 5–59.

Breymann W., Zumbach G., Dacorogna M. M., and Müller U. A., 2000, *Dynamical deseasonalization in otc and localized exchange-traded markets,* Internal document WAB.2000-01-31, Olsen & Associates, Seefeldstrasse 233, 8008 Zürich, Switzerland.

Dacorogna M. M., Gauvreau C. L., Müller U. A., Olsen R. B., and Pictet O. V., 1996, *Changing time scale for short-term forecasting in financial markets,* Journal of Forecasting, **15**(3), 203–227.

Davidson R. and MacKinnon J. G., 1993, *Estimation and Inference in Econometrics,* Oxford University Press, Oxford, England.

Drost F. and Nijman T., 1993, *Temporal aggregation of garch processes,* Econometrica, **61**, 909–927.

Engle R. F., 1982, *Autoregressive conditional heteroskedasticity with estimates of the variance of U. K. inflation,* Econometrica, **50**, 987–1008.

Engle R. F. and Bollerslev T., 1986, *Modelling the persistence of conditional variances,* Econometric Reviews, **5**, 1–50.

Guillaume D. M., Dacorogna M. M., and Pictet O. V., 1994, *On the intra-daily performance of garch processes,* In the proceedings of the First International Conference on High Frequency Data in Finance, Zürich.

Nelson D. B. and Cao C. Q., 1992, *Inequality constraints in the univariate garch model,* Journal of Business & Economic Statistics, **10**(2), 229–235.

Phillips P. and Park J. Y., 1988, *On the formulation of wald tests of nonlinear restrictions,* Econometrica, **56**(5), 1065–1083.

Press W. H., Teukolsky S. A., Vetterling W. T., and Flannery B. P., 1992, *Numerical Recipes in C. The Art of Scientific Computing,* Cambridge University Press, Cambridge.